DOCUMENT RESUME

ED 400 311 TM 025 683

AUTHOR Thompson, Tony D.; Pommerich, Mary
TITLE Examining the Sources and Effects of Local Dependence.
PUB DATE Apr 96
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; English; *Item Response Theory; Mathematical Models; Scoring; *Standardized Tests; Test Construction; *Test Reliability
IDENTIFIERS Conditional Independence; Dimensionality (Tests); *Item Dependence; *Local Independence (Tests); NOHARM Computer Program; Polytomous Scoring; Speededness (Tests); Three Parameter Model

ABSTRACT

Conditional item independence, also known as local independence, is necessary for the accurate estimation of item parameters within item response theory (IRT). Given that the condition of local independence will be violated to at least some degree when unidimensional models are used to represent multidimensional data, it is important to study the robustness of the unidimensional IRT models under these conditions. This paper identifies the sources and effects of local dependence for a national standardized achievement English test. The G-squared index and NOHARM (C. Fraser, 1988) residual analysis were selected as the most appropriate measures of evaluating the sources of local dependence. Results of the two analyses show that the last 20 items of the English test contain local dependence, which is primarily a speededness effect. It is also apparent that the item parameter estimates of the three-parameter model were affected by this dependence. Results are similar to those of previous research in that local dependence was found to affect estimated information and reliability when the test was scored polytomously. (Contains 6 tables, 6 figures, and 20 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 400 311

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
Tony D. Thompson

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

Examining the Sources and Effects of Local Dependence

Tony D. Thompson
Mary Pommerich

American College Testing

Paper presented at the annual meeting of the American Educational Research Association, New York, NY, April, 1996.

The authors thank Wen-Hung Chen, Tim Davey, Yaowen Hsu, Tim Miller, Alan Nicewander, Mark Reckase, and Jim Sconing for helpful discussions about the paper.

TM025683

BEST COPY AVAILABLE

Achievement tests tend to be multidimensional by nature (i.e., covering multiple content areas within an overall subject area). Nevertheless, unidimensional item response theory (IRT) models are commonly used to model the data. McDonald and Mok (1995) explain that misspecifying the dimensionality of data must result in item responses that are conditionally dependent, due to using the wrong conditioning variable(s). Conditional item independence, also known as local independence, is necessary for accurate estimation of item parameters within IRT. Given that the condition of local independence will be violated to at least some degree when unidimensional models are used to represent multidimensional data, it is important to study the robustness of the unidimensional IRT models under these conditions.

Recently, we have begun to conduct a series of research studies to identify both the sources and the effects of local dependence on two national standardized achievement examinations, each containing four subject matter tests. In this paper, we identify the sources and effects of local dependence for one of these tests.

Background

Let us begin our discussion of local dependence by first defining local independence under dichotomously scored models. Local independence is defined in the unidimensional case by the equation

$$P(\underline{U} = \underline{u}|\theta) = \prod_{j=1}^{n} P_j(U_j = u_j|\theta), \tag{1}$$

where $\underline{U}$ represents a response string of n binary variables, $P(\bullet)$ represents probability and θ is the person parameter representing latent ability or proficiency. This equation represents strict statistical conditional independence. A weaker condition can be written as

$$\rho(U_j, U_k|\theta) = 0, \quad j \neq k, \tag{2}$$

where ρ represents correlation. McDonald (1981) refers to the first equation as the *strong principle of local independence* and the second equation as the *weak principle of local independence*. Because this paper deals only with local dependence between item pairs, when we discuss local independence, it will be with reference to McDonald's weak principle of local independence.

Local dependence can stem from many sources, including content factors, speededness, passage-based items, fatigue, and practice effects (Yen, 1993). Identifying the source of local dependence is important because not all sources may be damaging to the test development process. For example, most achievement tests are multidimensional, representing multiple content areas, yet they report a single unidimensional score. Local dependence due to content factors, however, will not pose a major problem if test specifications tightly control the content contained across test forms. Although each form of a test may be multidimensional, it is reasonable to assume that stringent test specifications will ensure that the unidimensional reference composite (as defined by Wang, 1986) is the same across each form.

Other sources of local dependence such as speededness or passage-based items will be detrimental to the test development process, if the degree of local dependence is large. Oshima (1994) conducted a simulation study wherein speededness was defined as examinees not having time to respond to items at the end of the test. The study demonstrated that the *a* and *b* parameter estimates for items at the end of the test were inflated, with the most serious inflation occurring for the *a* parameter estimates. The inflation of the item parameter estimates could have serious implications for testing programs that routinely pretest items. If the pretest items are given at the end of a speeded exam, they will tend to have inflated parameter estimates, and these items will behave quite differently if placed at the beginning, rather than the end, of the operational form. In the case of a passage-based test, researchers have indicated that if the within-passage dependencies are ignored, the estimates of test reliability, test information, and item discrimination will be artificially inflated (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Thissen & Wainer, 1996; Yen, 1993).

As one way of coping with local dependence in passage-based tests, Wainer and Kiely (1987) introduced the notion of a "testlet". A testlet is created by summing the items within a passage, essentially treating the test as a small number of *independent* polytomous items, rather than a large number of potentially *dependent* dichotomous items. This method prevents within passage dependencies from affecting the polytomous item parameter estimates, though it does not eliminate possible dependencies across passages. Previous studies (Sireci, et al., 1991; Thissen, et al., 1989) have shown less information for dichotomous items grouped into testlets, and have assumed that the difference in information is due to local dependence inflating information estimates in the dichotomous case. Yen (1993) noted that even grouping locally *independent* items into testlets will result in a loss of information due to collapsing the dichotomous response pattern into testlets. Yen concluded, however, that the drop in information observed in her study was due primarily to local dependence.

In a previous study (Thompson & Pommerich, 1995), we evaluated a number of methods for detecting local dependence and identified sources of dependence on a wide variety of tests. In the current study, we expanded on our previous work to explore further the sources of local dependence. In addition, we employed the testlet approach to evaluate the effect of local dependence on reliability and information. In this study, we focused on one test from a national standardized examination that clearly displayed inter-item dependencies. The test used was a 75-item multiple choice English test, representing six content areas (e.g., punctuation, grammar, sentence structure) over five passages (14 items, 15 items, 15 items, 16 items, 15 items, respectively). Unlike a reading comprehension test where the items would follow each passage, the items on the English test were embedded within passages, so that examinees answered items as they were reading the passage. A random sample of 10,000 examinees was selected for the analyses.

Identifying the Sources of Local Dependence

A large number of methods exist that purport to measure local dependence. We previously applied five of these methods to two national standardized achievement batteries, each consisting of four different subject tests, to determine whether they gave consistent results (Thompson &

Pommerich, 1995). The methods considered included a test for unidimensionality, DIMTEST (Stout, Nandakumar, Junker, Chang, & Steidinger, 1991); two dependency measures based on item pairs, Q_3 (Yen, 1984) and G^2 (Chen & Thissen, in press); and two multidimensional IRT calibration programs, TESTFACT (Wilson, Wood, & Gibbons, 1987) and NOHARM (Fraser, 1988). These methods, in general, showed agreement in identifying the sources of local dependence. The sources we identified included speededness, content factors, and passages; for some tests, no appreciable dependence was noted. For our current study assessing the English test, we found that G^2 and the NOHARM residual analysis were the most useful methods for evaluating the sources of local dependence.

The G^2 Index

The G^2 index is designed to detect differences between observed and expected item contingency tables. For each pair of items, the observed and expected frequencies can be represented with 2×2 tables:

		Item j: 0	Item j: 1
Item i	0	O_{11}	O_{12}
	1	O_{21}	O_{22}

		Item j: 0	Item j: 1
Item i	0	E_{11}	E_{12}
	1	E_{21}	E_{22}

G^2 is defined as

$$G^2 = -2\sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij} \ln\left(\frac{E_{ij}}{O_{ij}}\right), \tag{3}$$

where the expected frequencies were obtained from a BILOG (Mislevy & Bock, 1990) calibration of the three-parameter logistic (3PL) IRT model (Lord, 1980). G^2 indices were computed using a program by Chen (1993). According to Chen and Thissen (in press), the G^2 distribution is approximately distributed as $\chi^2(1)$, and thus the expected value of the G^2 under the null distribution is approximately 1.0. W.-H. Chen (personal communication, January 9, 1996) has recommended that item pairs with G^2 values greater than 10.0 be flagged for possible local dependence.

We grouped the G^2 indices by passage and by content area and summarized over groupings; the results for passages are presented in Table 1. Table 1 contains the mean, standard deviation, minimum, median, and maximum G^2 values for all item pairs within and across passages. The table clearly indicates large dependencies within the last passage of the test; in every grouping, there were item pairs with large G^2 indices. Since the mean was greatly influenced by the

maximum values, the median appears to provide a better summary of the degree of local dependence within a grouping.

By visual inspection of the G^2 values at the item pair level, we determined that the largest local dependence occurred in mainly the last 20 items of the test. It appears that the local dependence is more attributable to a speededness effect, rather than a passage effect. By speededness, we mean that some examinees may not have had enough time to reach the items at the end of the test, or may have responded randomly to those items, or they may have rushed through the final questions in order to finish. To search for other possible sources of local dependence, we removed the last 20 items, recalibrated the item parameters, and recomputed the G^2 indices. For the item pairs summarized within and across passages, the median values gave no evidence for local dependence. The results for the item pairs summarized within and across content areas are given in Table 2. The median values suggest a possible dependence effect for items in the content area PUN (punctuation). When examining this summary, it should be noted that it is difficult to distinguish passage effects from content effects when the content is distributed across passages.

It was surprising to note that the mean G^2 values in Table 1 were much greater than 1.0 for all groupings (including across passages), and that the maximum values were so large. This suggested that our large sample may have overly influenced the G^2 values. When the G^2 indices were recomputed for the 75-item test using a sample size of 2000, the values were much smaller for each item pair; the results for item pairs summarized within and across passages are presented in Table 3. This analysis did not change our conclusions regarding the sources of local dependence, but it made us more wary of indiscriminately interpreting the G^2 values obtained from large samples.

The NOHARM Residuals

We performed 1-, 2-, 3-, 4-, and 5-dimensional exploratory multidimensional IRT calibrations (N=10,000), and examined the residuals of the item covariances after fitting each model. The NOHARM residuals for item pairs were interpreted in a similar manner as the G^2 indices. We grouped the residuals by passage and by content area and summarized over groupings; an advantage of the NOHARM analyses was that they seemed to be relatively unaffected by sample size. The passage summaries for the 1- and 2-dimensional solutions are presented in Tables 4-5. Although the residuals for the 1-dimensional solution appear small, the mean and median residuals were three to four times larger for Passage 5 items than for the other within passage residuals. This pattern suggests possible local dependence within Passage 5 items. The 2-dimensional solution appears to eliminate the local dependence within Passage 5. When the residuals for the 3-, 4-, and 5-dimensional NOHARM solutions were grouped by passage, the results were similar to those for the 2-dimensional solution. Because the factor loadings for the 2-dimensional solution show that the last 19 or 20 items loaded on the second factor, the primary dependence was determined to be the speededness effect.

Assessment of the G^2 indices and NOHARM residuals for the English test both led to the conclusion that the most prominent source of dependence was due to the speededness effect

noted in approximately the last 20 items[1]. A number of individual item pairs throughout the test displayed large G^2 values, but the only other systematic effects we identified were for punctuation items. Although many item pairs were identified as being dependent, it is difficult to know by these analyses alone if the dependencies detected would be large enough to affect a unidimensional calibration. Further, the G^2 results were apparently influenced by our large samples, and the NOHARM residuals appeared small, although Passage 5 contained relatively larger values. Identifying the existence and sources of local dependence is not sufficient to preclude the use of a unidimensional IRT model; it is necessary to directly examine the effect of the local dependence on test characteristics.

Evaluating The Effects of Local Dependence

Having detected local dependence, and identified the primary source for the English test, we examined the effect of local dependence on test information and test reliability. We computed information and reliability under two different scoring methods: (1) treating all items as dichotomous and (2) forming five testlets by summing items within passages. The second scoring method is the logical grouping that would follow from the testlet literature because the items within a passage are linked. Nevertheless, results from the detection of local dependence suggest that the grouping of passages 1-3 into polytomous items may not be necessary, as most of the local dependence seemed to reside in the last 20 items.

Both scoring methods were analyzed using MULTILOG (Thissen, 1991). When all items were treated as dichotomous, the 3PL IRT model was used to calibrate item and person parameters:

$$P(\theta) = c + \left(\frac{1-c}{1+e^{-1.7a(\theta-b)}} \right). \tag{4}$$

When the items were grouped into five testlets, Bock's (1972) nominal model was used. Under this model, the probability that a subject with latent ability θ will respond in category k ($k = 0, 1, \ldots, m_j$) on item j is given by

$$\Psi_{jk}(\theta) = \frac{\exp[a_{jk}\theta + c_{jk}]}{\sum_{h=0}^{m_j} \exp[a_{jh}\theta + c_{jh}]}, \tag{5}$$

where c_{jh} and a_{jh} ($h = 0, 1, \ldots, m_j$) are item parameters associated with category h of item j.

Calibration for the polytomous scoring method was also performed using a constrained version of the nominal model, where the a parameters were constrained to be equal for all categories within an item. A likelihood ratio χ^2 statistic (Thissen, et al., 1989) showed that the fit of the nominal model was superior to the fit of the constrained version, and therefore we report the results based on the nominal model.

---

[1] Preliminary G^2 and NOHARM analyses were performed for another form of the English test, administered in a different year. This form also demonstrated the speededness effect noted in the previous analysis. This provided some validation of our initial findings, but we have not reached the point where these results can be generalized across many forms of this test.

For each scoring method, test information estimates and reliability estimates were computed using standard IRT methods. Under the assumption of local independence, test information may be obtained as the sum of the item information functions. Note that this assumption is suspect for our test. Item information under the dichotomous model was estimated as in Lord (1980). Item information within the polytomous model was estimated as in Bock (1972). To estimate reliability, we followed Sireci, et al. (1991) in computing marginal reliability; so-called because it is derived from the marginal error variance of $\hat{\theta}$. Using the formulas given in their paper, marginal reliability $(\bar{\rho})$ was computed as

$$\bar{\rho} = \frac{\sigma_\theta^2 - \bar{\sigma}_e^2}{\sigma_\theta^2}, \tag{6}$$

where

$$\bar{\sigma}_e^2 = \frac{1}{\int_{-\infty}^{\infty} I(\theta) g(\theta) d\theta}, \tag{7}$$

where $I(\theta)$ is information. Our computations of information and reliability estimates were checked for accuracy against those given in MULTILOG.

When grouping locally *independent* dichotomously scored items into testlets, the collapsing of the dichotomous response patterns into a single score for each passage results in a loss of information and reliability. When items are locally *dependent*, we expect the dichotomously scored items to provide an inflated estimate of information and reliability. In this case, the difference in information between the dichotomous and polytomous scoring methods will be even greater. Thus, it is necessary to distinguish how much difference in information is due to the scoring method and how much is due to the existence of local dependence.

To make this distinction, we simulated data in order to obtain locally independent observations. The item parameters from the 75-item MULTILOG 3PL calibration were used as the true population parameters to generate the responses of 10,000 examinees under a N(0,1) latent ability distribution. The simulated examinee responses differed from the real examinee responses only in that they were generated from a model in which local independence held, whereas the real data presumably contained dependencies in the last 20 items. Thus, the test information functions for the real and simulated data should be approximately equal for dichotomous scoring. For the polytomous scoring condition, however, we expect the information function for the real data to be less than that for the simulated data, due to a reduction in information caused by the inter-item dependencies in the real data. If, on the other hand, there is little difference in information between the polytomous models for the real and simulated data, we would conclude that the effects of local dependence in the real data are of minor consequence.

An Aside

In order to compare test information and reliability across different scoring methods, it is necessary that the scoring methods have the same underlying latent ability metric. In discussions with our colleagues, some expressed the opinion that the results based on the 3PL and nominal

models must be rescaled to a common metric, while others questioned whether they could be compared at all. Because MULTILOG's underlying latent ability distribution is $N(0,1)$, regardless of scoring method, and because both methods are based on the same data, we believe that we can compare information and reliability without rescaling.

To address the concerns of our colleagues, we compared the latent ability metrics from the two scoring methods (75 dichotomous items versus 5 polytomous items). The correlation between the $\hat{\theta}$ s from the two models was .984, each distribution having a mean of approximately 0.0 and a variance of approximately 1.0. Although the empirical distribution of $\hat{\theta}$ should not be taken as being equivalent to the distribution of θ, these results demonstrate that the underlying latent ability distributions are probably similar, and that the models yielded very similar estimates of ability. In addition, comparing the test characteristic curves for the two models showed differences only at the extreme lower tail of the ability distribution ($\hat{\theta} \leq -2.5$), which is probably due to the models having different lower asymptotes. Out of curiosity, we conducted a true score equating using the test characteristic curves, and found that only nontrivial differences in information occurred, again at the extreme lower tail of the ability distribution ($\hat{\theta} \leq -2.5$). As our conclusions in this study would be unaltered with rescaling, the results we report will not be rescaled.

Results

The plot of estimated test information functions for the dichotomous and polytomous models is given in Figure 1, for both the simulated and real data cases. As predicted, the test information function for the 75 dichotomous items from the simulated data is nearly identical to the test information function for the 75 dichotomous items from the real data. This shows that the item parameter estimates of the real and simulated data are similar under the dichotomous case. The test information function for the 5 polytomous items in the simulated data shows less information than when the test was scored as 75 dichotomous items; this is due to the polytomous scoring. The test information for the 5 polytomous items in the real data is substantially less than the test information for the simulated polytomous items. This shows that one would obtain substantially less information scoring the real data polytomously than one would predict from the 3PL model, because the 3PL item parameter and information estimates are inflated.

To explore further where the difference in estimated information across the simulated and real polytomous models occurred, we plotted the information functions for each polytomous item. The plots are given in Figures 2-6. Because we found local dependence primarily in the last 20 items, we expected that the information functions for the first three polytomous items would show little differences across the simulated and real data; this expectation is confirmed by Figures 2-4. The fourth polytomous item (Figure 5) clearly shows some difference between the simulated and real information functions, with the simulated data showing more information primarily in the lower half of the ability distribution. We hypothesize this difference occurred because the last few items of the passage displayed local dependence due to the speededness effect.

The last polytomous item (Figure 6) demonstrates the substantial effect of local dependence in the real data. A large difference in estimated information exists between the real and simulated polytomous item nearly all throughout the ability distribution. This difference can be explained as follows: In the simulated data, the dichotomous items within this passage are highly discriminating and locally independent; this allows the polytomous item to discriminate well among examinees. In the real data, the dichotomous items have equally high (albeit incorrect) estimates of discrimination and information, but because of the local dependence in the data, there is less information in the item responses for the polytomous item to use in discriminating among examinees.

The marginal reliability estimates for the individual polytomous items in the real and simulated data cases are given in Table 6. This table also gives the test reliabilities for the 5 polytomously scored items and the 75 dichotomously scored items. The trends in the reliabilities mirror the trends in the information functions. The first three passages show similar reliabilities across the real and simulated data. Differences appear in the fourth and fifth passages. Particularly striking is the difference in reliability for polytomous item 5; reliability is .72 for the locally independent simulated data but only .28 for the locally dependent real data. The difference in test reliability is much smaller (.90 versus .93) between the real and simulated data for the polytomously scored items. Despite the low reliability observed in the last passage for the polytomous model in the real data, the overall reliability for the test is still respectable. Note that there was a small loss in test reliability for the simulated data when the responses were scored as 5 polytomous items rather than 75 dichotomous items, although with rounding reliability was .93 in both cases.

Conclusions

The results of our NOHARM and G^2 analyses showed that the last 20 items of the English test contained local dependence. Further, from our evaluation of the effects of local dependence, it can be ascertained that the item parameter estimates of the 3PL model were affected by this dependence. This was evident when we examined estimated information and reliability within the last two polytomous items; most notably in the last item. Even though the locally dependent (real) data had similar 3PL item parameter estimates to the locally independent (simulated) data for the last passage, the amount of estimated information obtained by the polytomous models for the last passage was dramatically different. The overall estimated reliability of the polytomously scored test for the real data was still respectable, but the contribution of the last passage to the information obtained by the test was vastly overrated by the dichotomous model.

Our study found results similar to previous studies given in the literature in that local dependence was found to substantially affect estimated information and reliability when the test was scored polytomously. Like Yen (1993), we provided evidence that the drop in information was not due to simply the method of scoring, although our methodology differed from hers. A major difference between our study and previous ones was the source of local dependence. Passages were found to cause little dependence in our study, no doubt because in our test the items were embedded within the passages, unlike traditional reading passages. Speededness, on the other hand, was found to be a major source of dependence. We were able to adopt the testlet

methodology to help us examine the effect speededness had on estimated information and reliability.

A recommendation to a researcher modeling data of this nature might be to use a mixed model. Treat the first 55 items as dichotomous items and the last 20 items as a testlet to avoid the effects that local dependence would have on the item parameter estimates. This would give a more accurate representation of the test information and reliability than either the 75 dichotomous items case or the 5 polytomous items case. Note that this approach in no way would eliminate the speededness effect in the data; it simply reduces the effect of local dependence on the item parameters. For a test developer desiring to treat all items as dichotomous and calibrate all items accurately, we can only recommend that more time be given to examinees. Extended testing time would reduce the need for some examinees to hurry through the latter portion of the test, although it is likely that some examinees would still exercise poor time management, no matter how long the time limit.

We would be reluctant to accept the 3PL item parameter estimates from the last passage at face value. But we do not yet feel that we have determined just how poor those estimates might be. We have shown that dependencies in the data affect the information provided by a polytomous model, but we would like more direct evidence of their effect on the 3PL model. We are planning a new study to evaluate the deterioration in item parameter estimates due to local dependence. We plan to use a 6-dimensional NOHARM solution to accurately represent dependencies in the real data. After calibrating dichotomous and polytomous models on simulated examinees, we will obtain through simulation empirically derived estimates of information for various θ values along the ability scale. Comparing these with the information predicted by the models assuming locally independent data should give us a clear picture of the inflation of information that arises from using a dichotomously scored model with locally dependent data.

References

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37*, 29-51.

Chen, W.-H. (1993). *IRT_LD: A computer program for the detection of pairwise local dependence between test items*. Research Memorandum 93-2. Chapel Hill, NC: L.L. Thurstone Laboratory, University of North Carolina at Chapel Hill.

Chen, W.-H., & Thissen, D. (in press). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*.

Fraser, C. (1988). *NOHARM II: A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory*. Armidale, N.S.W.: University of New England, Centre for Behavioral Studies.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.

McDonald, R., & Mok, M.M. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30*, 23-40.

Mislevy, R.J. & Bock, R.D. (1990). BILOG3: *Item analysis and test scoring with binary logistic models*. Chicago: Scientific Software.

Oshima, T.C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*, 200-219.

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247.

Stout, W., Nandakumar, R., Junker, B., Chang, H., & Steidinger, D. (1991). *DIMTEST and TESTSIM*, programs for dimensionality testing and test simulation. University of Illinois at Urbana-Champaign, Department of Statistics.

Thissen, D. (1991). *Multilog user's guide*. Chicago: Scientific Software.

Thissen, D., Steinberg, L., & Mooney, J.A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*, 247-260.

Thompson, T.D., & Pommerich, M. (1995, June). *Evaluating testlets as a strategy for managing local item dependence.* Paper presented at the Meeting of the Psychometric Society, Minneapolis, MN.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15(1)*, 22-29.

Wang, M. M. (1986, April). Fitting a unidimensional model to multidimensional item response data (ONR Rep. 0422860). Iowa City IA: University of Iowa.

Wilson, D., Wood, R., & Gibbons, R.D. (1987). *TESTFACT: Test scoring, item statistics, and factor analysis.* Mooresville, IN: Scientific Software, Inc.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Table 1. G^2 Indices for 75 Dichotomous Items Summarized Over Passage Groupings (N=10,000)

Grouping	Passage	Mean	SD	Minimum	Median	Maximum
Within Passages	1	4.24	8.52	0.00	1.77	74.48
	2	6.59	11.76	0.05	2.38	97.94
	3	3.35	4.16	0.02	1.61	16.06
	4	8.29	17.37	0.06	2.94	120.79
	5	59.43	111.59	0.18	20.44	789.80
Between Passages	1 and 2	6.94	26.32	0.01	2.23	309.16
	1 and 3	3.26	6.67	0.01	1.08	46.17
	1 and 4	3.13	5.43	0.00	1.27	51.04
	1 and 5	6.04	6.22	0.03	3.92	30.99
	2 and 3	4.40	7.74	0.02	1.69	62.38
	2 and 4	5.20	15.83	0.00	1.97	220.26
	2 and 5	10.30	11.96	0.04	7.08	75.84
	3 and 4	3.83	6.62	0.00	1.51	55.26
	3 and 5	5.47	5.97	0.01	3.40	30.87
	4 and 5	8.67	17.65	0.06	3.30	156.04
Total	All	7.77	27.11	0.00	2.44	789.80

Table 2. G^2 Indices for the First 55 Dichotomous Items Summarized Over Content Area Groupings (N=10,000)

Grouping	Content	Mean	SD	Minimum	Median	Maximum
Within Content Area	SST	3.63	5.27	0.01	1.99	29.59
	ORG	1.83	2.40	0.01	0.83	9.70
	BGU	2.90	4.18	0.04	1.18	15.97
	PUN	35.96	66.87	0.05	7.02	270.62
	STY	2.80	3.29	0.01	1.29	11.82
	STR	4.53	6.92	0.02	2.15	32.41
Between Content Area	1 and 2	2.70	4.21	0.01	1.22	37.59
	1 and 3	2.50	5.84	0.02	0.98	63.65
	1 and 4	5.92	11.51	0.04	2.07	84.04
	1 and 5	3.94	5.63	0.03	1.41	26.37
	1 and 6	2.35	3.06	0.02	0.95	18.88
	2 and 3	1.27	2.09	0.00	0.44	15.72
	2 and 4	2.70	3.48	0.00	1.08	13.14
	2 and 5	3.53	4.85	0.00	1.66	29.26
	2 and 6	1.47	2.39	0.00	0.62	14.71
	3 and 4	1.60	1.79	0.02	0.98	7.55
	3 and 5	2.13	2.77	0.02	0.72	11.93
	3 and 6	1.20	1.18	0.02	0.75	4.13
	4 and 5	6.52	16.13	0.02	2.31	102.70
	4 and 6	2.24	3.62	0.00	0.95	19.14
	5 and 6	1.66	1.41	0.01	1.43	4.98
Total	All	3.51	11.82	0.00	1.12	270.62

Table 3. G^2 Indices for the 75 Dichotomous Items Summarized Over Passage Groupings (N=2000)

Grouping	Passage	Mean	SD	Minimum	Median	Maximum
Within Passages	1	1.63	2.28	0.01	0.83	13.40
	2	2.35	3.37	0.02	1.21	24.96
	3	1.31	1.85	0.02	0.70	8.67
	4	3.46	6.26	0.02	1.04	40.00
	5	18.37	33.75	0.00	8.00	250.64
Between Passages	1 and 2	2.44	6.66	0.00	0.84	81.33
	1 and 3	1.61	2.30	0.00	0.74	15.74
	1 and 4	1.27	1.65	0.01	0.55	8.83
	1 and 5	3.06	3.35	0.00	1.88	16.86
	2 and 3	1.68	2.55	0.01	0.79	20.70
	2 and 4	1.77	3.76	0.00	0.76	40.29
	2 and 5	4.05	4.01	0.00	2.59	25.60
	3 and 4	1.56	2.39	0.01	0.66	22.31
	3 and 5	2.13	2.37	0.01	1.20	15.19
	4 and 5	2.51	4.61	0.00	1.13	37.79
Total	All	2.82	8.17	0.00	1.02	250.64

Table 4. NOHARM residuals from 1-Dimensional Solution for 75 Dichotomous Items Summarized Over Passage Groupings (N=10,000).

Grouping	Passage	Mean	SD	Minimum	Median	Maximum
Within Passages	1	-.002	.003	-.015	-.002	.005
	2	-.003	.004	-.017	-.003	.008
	3	-.002	.003	-.011	-.002	.005
	4	-.002	.005	-.020	-.002	.008
	5	-.011	.011	-.059	-.009	.006
Between Passages	1 and 2	-.003	.005	-.036	-.002	.007
	1 and 3	-.002	.003	-.013	-.001	.007
	1 and 4	.001	.003	-.010	.008	.010
	1 and 5	.004	.004	-.007	.004	.014
	2 and 3	-.002	.004	-.016	-.002	.015
	2 and 4	.001	.005	-.034	.001	.012
	2 and 5	.005	.005	-.021	.005	.020
	3 and 4	.000	.004	-.016	.000	.016
	3 and 5	.004	.004	-.008	.005	.013
	4 and 5	.001	.006	-.033	.002	.016
Total	All	.000	.006	-.059	.000	.020

Table 5. NOHARM residuals from 2-Dimensional Solution for 75 Dichotomous Items Summarized Over Passage Groupings (N=10,000).

Grouping	Passage	Mean	SD	Minimum	Median	Maximum
Within Passages	1	-.000	.003	-.013	-.000	.007
	2	-.000	.004	-.016	.000	.011
	3	-.001	.003	-.008	-.000	.006
	4	-.002	.005	-.017	-.002	.008
	5	-.001	.007	-.023	.001	.013
Between Passages	1 and 2	-.000	.004	-.027	.000	.009
	1 and 3	.000	.003	-.011	.000	.010
	1 and 4	.001	.003	-.011	.001	.008
	1 and 5	-.001	.004	-.019	-.000	.007
	2 and 3	.000	.004	-.015	.000	.016
	2 and 4	.001	.004	-.033	.001	.010
	2 and 5	-.001	.004	-.021	-.000	.011
	3 and 4	.000	.004	-.015	.000	.016
	3 and 5	.000	.003	-.010	.001	.008
	4 and 5	.001	.005	-.033	.002	.011
Total	All	-.000	.004	-.033	.000	.016

Table 6. Marginal Reliability Estimates.

Grouping	Simulated Data	Real Data
Polytomous Item 1	.39	.41
Polytomous Item 2	.56	.52
Polytomous Item 3	.62	.63
Polytomous Item 4	.68	.58
Polytomous Item 5	.72	.28
5 Polytomous Items	.93	.90
75 Dichotomous Items	.93	.93

Figure 1. Estimated Test Information Functions for the Dichotomous and Polytomous Models, Real and Simulated Data.

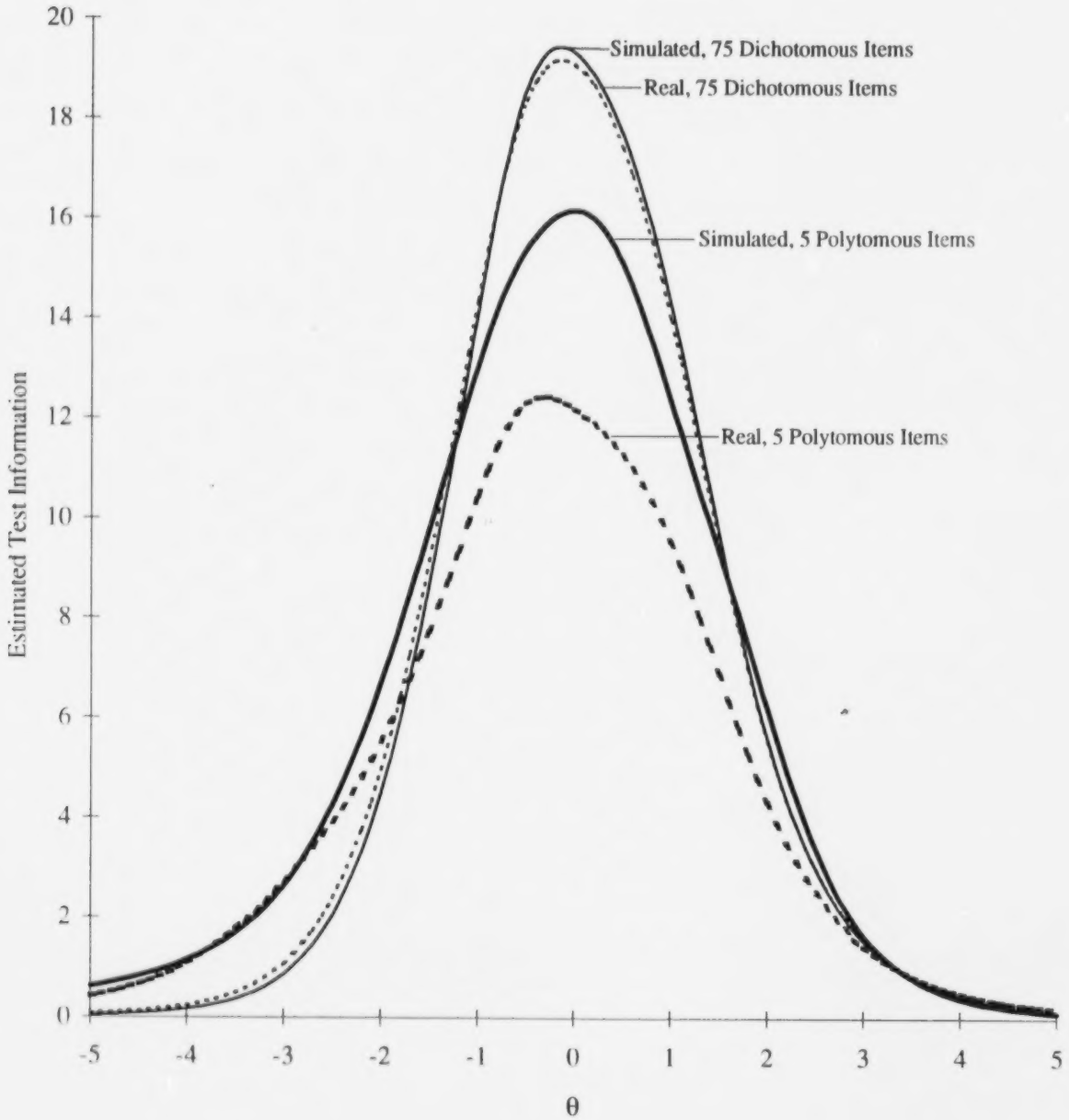

Figure 2. Estimated Item Information Functions for Item 1 of the Polytomous Models, Real and Simulated Data.

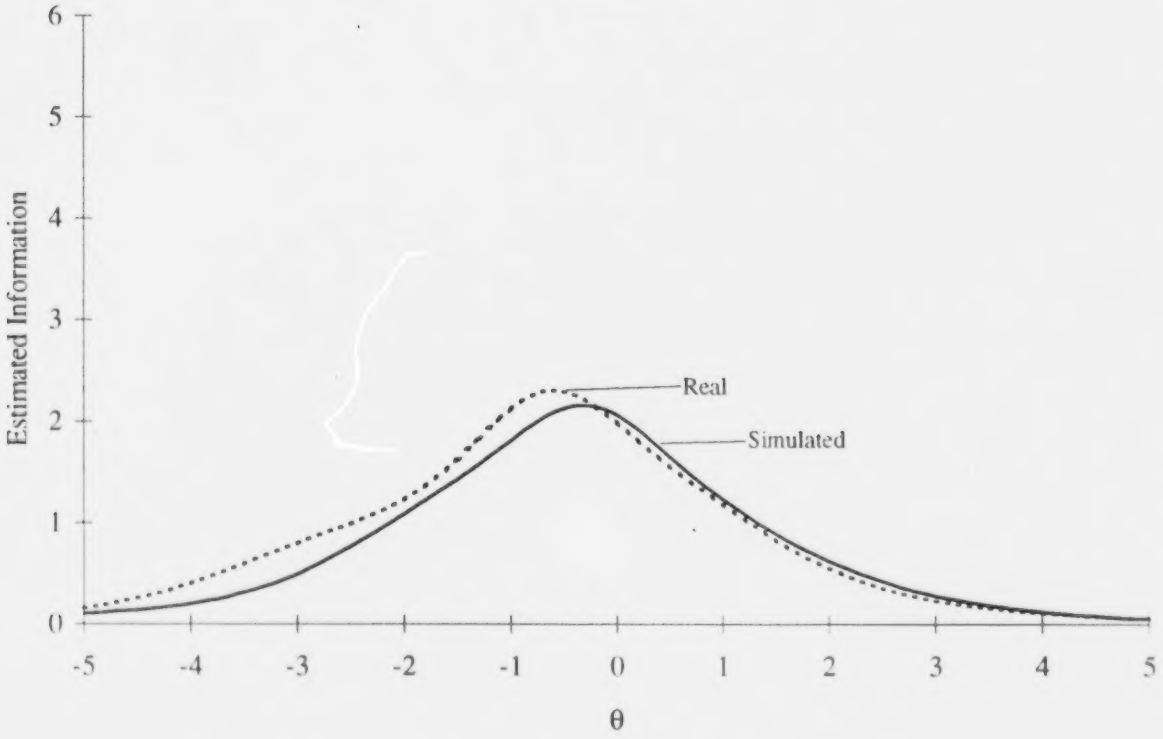

Figure 3. Estimated Item Information Functions for Item 2 of the Polytomous Models, Real and Simulated Data.

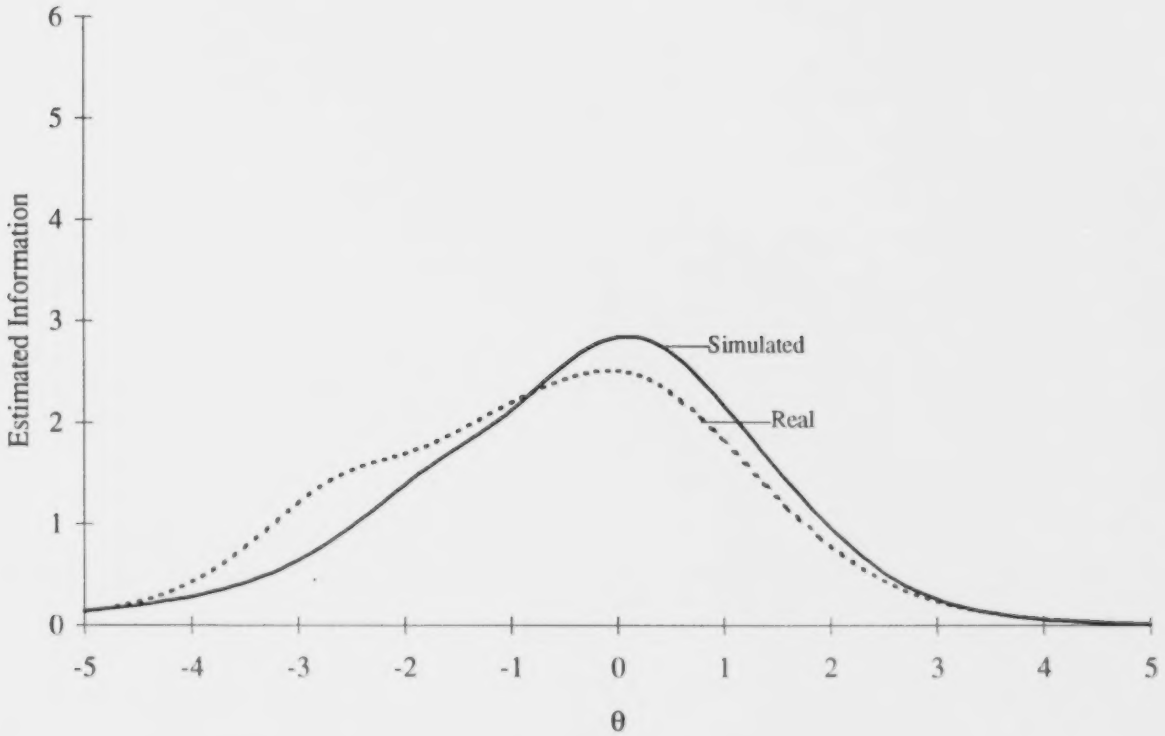

Figure 4. Estimated Item Information Functions for Item 3 of the Polytomous Models, Real and Simulated Data.

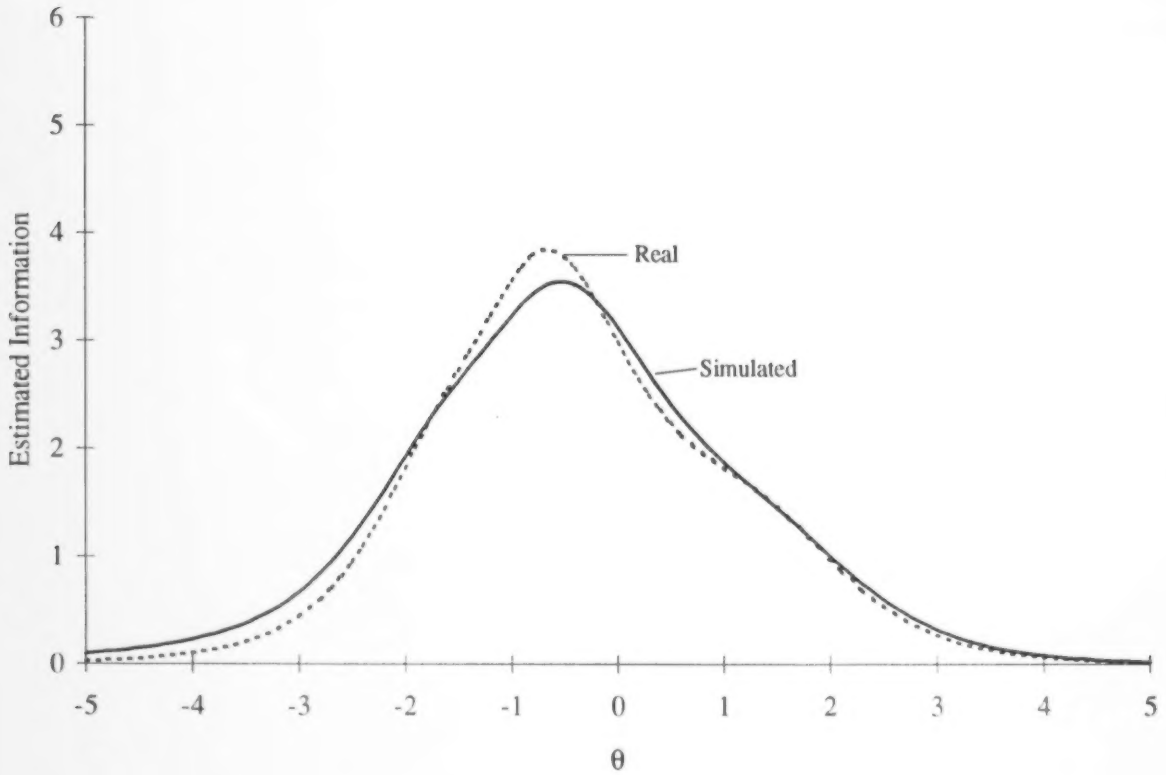

Figure 5. Estimated Item Information Functions for Item 4 of the Polytomous Models, Real and Simulated Data.

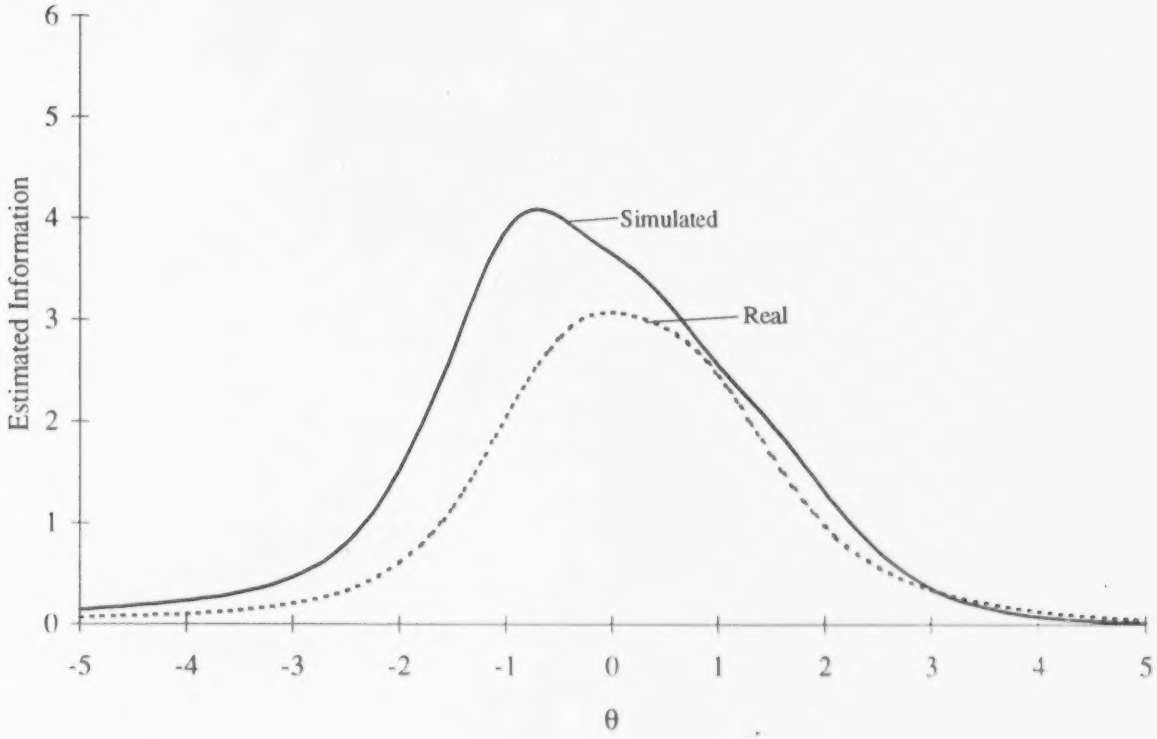

Figure 6. Estimated Item Information Functions for Item 5 of the Polytomous Models, Real and Simulated Data.